Research Paper | Engineering | E-ISSN No : 2454-9916 | Volume : 3 | Issue : 12 | Dec 2017

# A BRIEF OVERVIEW OF AUTOMATIC DOCUMENT SUMMARIZATION

Abhishek Sathe

Department of Computer Engineering, Sinhgad College of Engineering, Pune, India - 41.

## ABSTRACT

With the advent of the Internet, a large amount of data is being generated on a daily basis. The World Wide Web contains billions of documents, images and pictures and these documents are growing at an exponential rate. It is not practically possible to go through each and every document manually and find out the relevant information that might be required for a particular research. Document summarization is the process of shortening the document using a certain software, in order to create a summary of the software with only the important points from the original document. Document summarization can be done either for a single document or for multiple documents. There are various techniques for document summarization such as extractive and abstractive technqiues. In this paper, I briefly introduce the topic of automatic document summarization.

**KEYWORDS:** Document Summarization,Data Minin, Statistical Analysis, Abstractive Summary, Extracts.

## 1. INTRODUCTION:

Worldwide Information is becoming the key and most of the information is online as the world increasingly embraces technology and increasing dependence on collaboration through social and mobility. The World Wide Web contains billions of documents, images and pictures, and these documents are growing at an exponential rate. The size of these document information can vary to a great extent, from concise to even thousand pages long. Such a large volume of information makes it difficult to search and find the exact or important information that is required for either a research or a study or any such purpose. Such concerns have spawned the interest in the development of automatic document summarization. Automatic document summarization is the process of shortening the document using software, in order to create the summary of the document with only the important points from the original document. As humans, we usually summarize by reading the document/image, understand and interpret and summarize it focusing on the main points and highlighting it. Since computers lack human knowledge and language capability, it makes Automatic Text Summarization a very challenging, difficult and non-trivial task.

Formally, automatic text summarization may be defined as [2] the task of producing a concise and fluent summary while preserving key information content and overall meaning. Microsoft Word's AutoSummarize is a simple example of text summarization.

Automatic summarization may be classified [2] as:

1) *Extractive summarization:* A process which summaries (also called extracts) are generated by concatenating several sentences taken exactly as they appear in the original text document.

2) *Abstractive summarization:* It is a technique of generating a summary from a text from its main ideas, not by copying most salient features from the text. The main information in the input document is expressed in the words of the summary author.

## 2. EARLY METHODS:

Historically, document summarization used sentences to find the summary of the entire document. One of the earliest work on automatic summarization [3] was of Luhn in the 1950s. In the 1950s document summarization [3] was carried out using salient features of sentences such as word and phrase frequency (Luhn,1958), position in the text (Baxendale,1958) and key phrases(Edmunson 1969). Many of the later published works focused on other domains, particularly on newswire data. Many approaches addressed the problem by building summarization systems depending upon the type of the summary required.

## 3. EXTRACTIVE SUMMARIZATION:

As stated before, extractive summarization techniques produce summaries by choosing a subset of the sentences from the original text document. The summary will contain some of the most important statements from the original document. Input type for an extractive summarization can be a single document or multiple documents. The extractive summarizations techniques can be easier to implement as compared to the abstractive summarization techniques, but the output produced by abstractive summary is more accurate [4]. Search engines are an example of extractive summarization technique. An example of how extractive summarization works in search engines can be seen from figure 1.

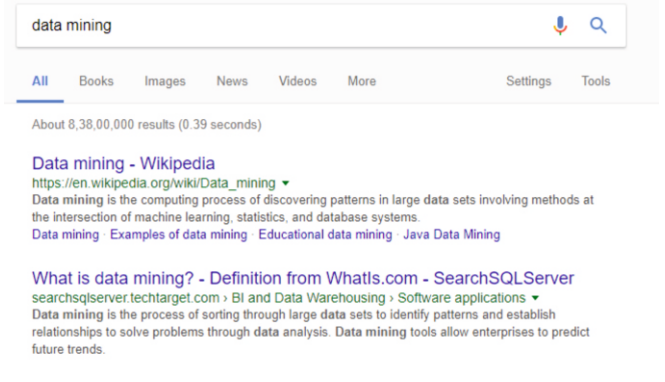


**Fig 1: Example of Extractive Summarization**

Extractive summarization involves a simple algorithmic approach in which each sentence in the document is analyzed for the presence or absence of certain salient features and based on these features it is decided whether to include them in the summary. The implication of sentence is based on linguistic and statistical features.

### 3.1 Extractive Summarization Techniques:

An extractive summarization method consists of select important sentences, paragraphs etc. from the original source document and concatenating them into a shorter form. These techniques aim at picking out the most relevant sentences in the document while also maintaining a low redundancy in the summary[1].

**A] Term Frequency-Inverse Document Frequency (TF-IDF) approach:**
A term frequency means how many times a word appears in a given document. The Inverse Document Frequency implies the number of times a word appears in a corpus of documents. TF-IDF calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in. It is a numerical statistic that is intended to demonstrate how important a word is in a given document. Words with a high TF-IDF value imply a strong relationship with the document they appear in. The TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. The TF-IDF weight consists of two terms [5] :

1) **Term-Frequency(TF):** This term implies the number of times the word t appears in the given document, divided by the total number of words in the document.

2) **Inverse Document Frequency(IDF):** This term measures how important a term is. It can be calculated as:

$$IDF_{(t)} = \log(N/N_t) \quad \ldots(1)$$

Where,
N = Total Number of documents

$N_t$ = Number of documents with t in it

The weight $WT_{(t)}$ is then calculated as:

$$WT_{(t)} = TF_{(t)} * IDF_{(t)} \quad \ldots(2)$$

**B] Fuzzy Logic Approach:**

Fuzzy Logic is a mathematical tool that is used for dealing with uncertainity, ambiguity or vagueness. Fuzzy logic approach is used to find whether a member belongs to a particular set. In summarization, fuzzy logic approach can be used to determine the degree of importance of the sentence to the final summary. The fuzzy logic system consists of four differenct components : Fuzzifier, Rule Base, Inference Engine and Defuzzifier. In Fuzzifier, the crisp input values are converted into linguistic values in the range of 0 to 1 for each feature. It then uses a membership function to decide fuzzy sets for input values. The membership function (MF) is used to assign a membership value between 0 and 1 in the input space. The most commonly used membership functions are Gaussian,Sigmoid,triangular,cubic polynomial curves etc. The Fuzzy Inference Engine is the process of formulating the mapping from a given input to an output using the Fuzzy rule base. Mamdani fuzzy inference engine is the most commonly used [4] , due to its simple nature of min-max operations. The Rule Base consists of a set of rules that are used for the decision making process. The Defuzzifier performs an operation that is exactly opposite to what the fuzzifier does,it converts the linguistic values obtained from the inference engine block into crisp values. The membership functions are used for representing the final sentence score. The built-in methods used by defuzzification process are centroid, bisector middle of maximum and smallest of minimum. In fuzzy logic method, all the sentences are ranked in a descending order depending upon their sentence score. A set of sentences with the highest score are then extracted as the final summary.

**C] Graph Based Approach:**

Graph methods are influenced by the Page Rank algorithm [6]. In graph based approach, the documents are represented as graphs where the vertices represent the sentences. The edges between the vertices indicate how similar two sentences are. A similarity measure is employed to find out how closely two sentences are related to each other. If the measure is greater than a certain threshold, then the sentences are closely related. TF-IDF method is most commonly used to determine the similarity measure. This graph representation results into two outcomes. First, the partitions (sub-graphs) indicate discrete topics covered in the original document. The second outcome is the identification of the most important sentences in the graph that are to be included in the final summary. Sentences in a sub graph that are connected to many other sentences have a high probability to be included in the final summary. This graph based approach can be used for single documents as well as for multiple documents.

**D] Neural Network Method:**

A neural network is a system of programs and data structures that approximates the operation of the human brain. A neural network generally involves a large number of processors operating in parallel, each with its own small sphere of knowledge and access to data in its local memory. The first phase of this process involves training the neural network with data that will include the type of sentences that have to be included in the final summary. The network first has to be trained with several test paragraphs where each sentence is to be identified whether it has to be included in the summary or not [3]. This has to be done by a human reader. The neural network thus 'learns' the patterns and features that are inherent in the sentences and identifies the sentences that have to be included in the final summary. The feature fusion phase eliminates the uncommon features among the sentences and collapses the effects on sentences of the common features. This phase then finalises features that must be included in the summary sentences by combining the features and finding similar patterns in the summary sentences. In the final phase, the network can be used as a tool to filter any sentences in a given paragraph and determine whether each sentence should be included in the summary or not.

**E] Machine Learning Method:**

This approach is based on Bayes' Theorem of Inverse Probability and is also called as the Probability approach [4]. This approach models summarization as a classification problem. Sentences are included or excluded from the summary by calculating their probabilities of relevance, using the Bayes theorem. Decision Trees, Naïve Bayes, Hidden Markov models are some of the most common machine learning techniques that are used for summarization [5].

**F] Latent Semantic Analysis:**

Latent Semantic Analysis (LSA) is a robust unsupervised technique for deriving an implicit representation of text semantics based on observed co-occurrence of words. In LSA, the sentence is represented as matrix, in which the row represents the word and the column represents the sentences. Each entry aij in the matrix corresponds the weight of the word. If the sentence contains the word, the weight is equal to TF*IDF value of the word. If the sentence does not contain the word, its weight is equal to zero. To represent the matrix as a product of three different vertices, standard techniques such as singular value decomposition (SVD) are applied to the matrix. From the decomposed matrices, the sentences are selected using various algorithms [5] such as Gong and Liu approach, Ozsoy approach etc. The advantage of using LSA vectors is that conceptual (or semantic) relations as represented in the human brain are automatically captured in the LSA. It has the ability to collect all trends and patterns from each of the sentence.

## 4. ABSTRACTIVE SUMMARIZATION

Abstractive sentence summarization generates shorter version of a given sentence while preserving the meaning of the original sentence. Abstractive models generate summaries from scratch without being constrained to reuse phrases from the original text. Extractive summarization techniques simply copy the text from the original document to form the summary. Abstractive summarization involves paraphrasing the original document. The system infers the meaning of the original document and the important points are restated in the final summary. Abstractive summarizers are harder to implement as compared to extractive summarizers, but they are more accurate [2].

## 5. CONCLUSIONS:

A large amount of information is being generated on a daily basis on the internet. It is quite impossible to manually summarize such volumes of data to obtain the important point. Therefore, automatic document summarizing is the need of the hour and fueling a huge amount of reseaarch.

In this paper, I have provided a glimse of the various extractive approach in document summarization. The two methods given for text summarization have pros and cons. The extraction technique is simple to implement but can produce ambiguous results. The abstractive summarizers are harder to implement but produce accurate results. Although it is not possible to explain all possible algorithms and techniques in this paper, we think that it provides a good insight into the field of automatic document summarization and describes the recent trends in this area.

## REFERENCES :


1. Vishal Gupta and Gurpreet Singh Lehlal, "A Survey of Text Summarization Extractive Techniques",Journal of Emerging Technologies in Web Intelligence, Vol. 2, No.3, August 2010.
2. Rasha Mohammed Badry, Ahmed Sharaf Eldwin, Doaa Saad Elzanfally,"Text Summarization within the Latent Semantic Analysis Framework: Comparative Study", Internation Journal of Computer Applications,Vol. 81,No 11.
3. Vijay Sonawane, Rakesh Salam,"Graph Based Approach for Multi Document Summarization",International Journal of Engineering and Computer Science",Vol. 4,No. 4,April 2015.
4. Mahak Gambhir, Vishal Gupta,"Recent Automatic Text Summarization Techniques: A Survey",Artificial Intelligence Review,Vol. 47,No. 1,January 2017.
5. J. Goldstein, M. Kantrowitz, V. Mittal, J. Carbonell, "Summarizing Text Documents: Sentence Selection and Evaluation Metrics",ACM-SIGIR,1999, pp 121-128.
6. A. Khan,N. Salim."A Review on Abstractive Summarization Methods",Journal of Theoretical and Applied Information Technology",Vol. 59,No. 1,pp. 64-72,2014.